DOCUMENT RESUME

ED 047 003 TM 000 371

AUTHOR Messick, Samuel; Anderson, Scarvia
TITLE Educational Testing, Individual Development, and Social Responsibility.
INSTITUTION Educational Testing Service, Princeton, N.J.
PUB DATE Nov 70
NOTE 24p.

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Cultural Differences, Culture Free Tests, Disadvantaged Youth, *Individual Development, Item Analysis, *Minority Groups, Predictive Validity, *Test Bias, Test Construction, *Testing, Test Interpretation, *Test Validity

ABSTRACT

Recent harsh criticisms that educational and psychological tests are unfair and inadequate measures of the capabilities of minority, poverty, and other educationally alienated groups are discussed. The authors suggest that there are two main issues, the first scientific, the second ethical: (1) Is a test a valid measure of the characteristics it purports to assess for particular types of individuals in particular circumstances; and (2) the whole question of test use, beginning with whether or not a test should be utilized for a specified purpose? Responsible standards exist for evaluating the adequacy and appropriateness of a test for a particular use, but they are not always applied. The adequacy of measurement and the question of bias, the appropriateness of test use and the question of fairness, the side effects of testing, the problems of misinterpretation and secondary use of tests, the ethics involved, and the social consequences of not testing are other important topics discussed and analyzed in some detail. (CK)

ED047003

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

# Educational Testing, Individual Development, and Social Responsibility

Samuel Messick and Scarvia Anderson
Educational Testing Service

Educational and psychological tests have been harshly criticized on a number of occasions recently on the grounds that they are unfair and inadequate measures of the capabilities of particular groups of individuals--especially those from minority and poverty backgrounds and others who for a variety of reasons are educationally alienated. Robert L. Williams and the Association of Black Psychologists, for example, have called for "a moratorium on the repeated abuse and misuse of the so-called conventional psychological tests" because they "are unfair and improperly classify Black children." Indeed, they demand "an immediate moratorium on all testing of Black people until more equitable tests are available" (Williams, 1970 a). Implicit in this indictment is the premise that most educational and psychological tests are intrinsically biased against minority/poverty individuals and that their very use with these groups is misuse.

Responses to these charges from the testing community usually insist that the blame is misplaced--that it is not tests per se that are at fault but rather the recurrent misuses of tests in particular applications. What is needed, they say, is not so much

TM 000 371

the development of more equitable tests as the elimination of unfair and inequitable testing practices.

A complicating feature of this interchange is that the reaction, although plausible and by and large correct, is not directly responsive to the charge. The charge questions the adequacy of most tests; the response admits the inadequacy of much testing practice. There are really two issues here that not only are separable but must be separated if we are to recognize that there are multiple sources of discontent and multiple courses of remedial action, each course considerably more constructive than a monolithic demand for the abolition of testing. One issue deals with the whole question of whether a test is any good--for particular types of individuals under particular circumstances--as a measure of the characteristics it purports to assess. The other issue deals with the question of test use, beginning with whether or not a test <u>should</u> be utilized for a specified purpose. The first question is a scientific one; it may be answered by appraising the test's psychometric properties, especially its construct validity. The second question is an ethical one; it must be answered by evaluating the potential consequences of the testing in terms of human values (Messick, 1965; Jackson & Messick, 1967). Both questions must be addressed whenever testing is considered.

In this paper we shall point out (a) that responsible standards exist for evaluating the adequacy and appropriateness of a test for

a particular use and (b) that for a variety of reasons these standards are not always applied, leaving considerable room for improvement if testing is to fulfill its potential as a positive force for promoting education, training, and opportunity. We shall emphasize the various possibilities for improvement--or the many potentialities of testing--rather than the shortcomings of the present state of the field, because we feel that the indictment of testing by Williams and others must ultimately be met in terms of improved test development, application, and interpretation. Furthermore, these improvements must take into account more than the single issue of improper measurement or classification that their charge implies.

Before considering the scientific basis for evaluating whether a test measures the same thing with the same fidelity in different racial and other population groups, we will first discuss the importance, in dealing with these basic issues of bias and validity, of expanding the typically restricted definition of "test" to a more general notion of "assessment" broadly conceived. We will then turn to the problem of evaluating the appropriateness of test use in terms of the potential social consequences of the testing, underscoring the need to take into account the possibility of positive and negative side effects on both the person being tested and the person doing the testing. Finally, in view of the seriousness of the recent call for a moratorium on the testing of Black people, we will point to some of the critical social consequences of not testing.

## Testing As Systematic Inquiry

The statements of Williams and other spokesmen for the Black community suggest that their concerns about tests are addressed almost entirely to those instruments associated with "intelligence," "IQ," and "verbal and quantitative aptitude." It is probably the case that such measures have been the ones most frequently misused, but this is partly because they have been the most frequently used. Many members of the educational establishment restrict their conception of "tests" to a similar narrow range, and this limited perspective has led both to the use of intelligence and aptitude tests in situations where they were unsuitable or unproductive and to the failure to seek and develop other means of assessing student performance, appreciation, knowledge, understanding, and judgment.

"Playing the Dozens" is a verbal jousting that depends on an imaginative derogation of the participants' backgrounds (see Williams, 1970a). Frequently a series of episodic references are cited as the verbal interplay is intensified. It would take very little doing to turn a Dozens "game" into a test. It already has some elements of a standard stimulus (an "enemy" who is to be destroyed through attacks on his parents), a circumscribed response system (verbal, oral, in quatrains), and a scoring system (as Brown, 1969, describes it, "...the winner was determined by the way they responded to what you said. If you fell all over each other laughing, then you knew you'd

scored..."). Whether or not it would be a "good" test would depend upon judgments about the importance of the content and purpose of the measurement and upon its properties in relation to other measures and performance criteria, balanced against the possible harm that such measurement might engender.

It is possible that good performance on such a test would be predictive of good performance on other verbal fluency measures and perhaps of leadership in the peer group. But it is doubtful that any educational program would take it upon itself to try to improve performance in playing the Dozens of those who earned low scores on the test.

Nevertheless, it is important to make the point that there are many qualities of student behavior that need to be assessed in order to identify talent and to initiate educational programs relevant to the needs of individuals. Furthermore, these qualities can be assessed in a context that is compatible with the student's previous experiences and thus does not introduce the irrelevant difficulty of "strangeness." This strangeness or the perceived irrelevance of the test to the life experiences of the examinee represents a kind of face invalidity, if you will, which poses a constant potential threat to the psychometric validity of the assessment in individual instances. But this strangeness is relative: its impact can be reduced by instruction and practice and, since it is not a necessary concomitant of the testing process, it

may even be avoided completely by sensitive test construction and administration. For example, a second grade teacher can "test" a child's social competencies by observing him in play with his peers, his vigor in normal activities like jumping, his understanding of money through "store" exchanges, his attitudes toward members of other ethnic groups through his choice of ethnically identified playthings, his listening comprehension ability through his reaction to television messages, his manipulative skills through toy assembly projects, his interest in his school work by his eagerness to get started in the mornings, aspects of his imagination through his art work, etc. Furthermore, tasks and observations of these types can readily be standardized and even "scaled" to the extent that the teacher can order or sort the children in her class in terms of their needs for special instruction or experience.

## The Adequacy of Measurement and the Question of Bias

The Association of Black Psychologists has charged that conventional tests improperly classify Black children. It is indeed true that individuals from minority and poverty backgrounds typically obtain lower scores on conventional tests than members of the White middle class, who dominate most norms groups. But it is important to inquire into the possible sources of this poorer performance, because some of the contributing factors can be counteracted. Let

us consider three of these sources.

1. The Test May Measure Different Things for Different Groups

It is possible for the same test to measure different attributes or processes in minority/poverty groups than it measures in White middle-class samples or for the same processes to be captured with a different degree of fidelity. If this is the case, then scores should certainly not be interpreted in the same way in both groups, nor should performance levels in one group be compared with those in the other as if they were on the same dimension. To discount the possibility that the same instrument measures different things in different groups, it is necessary to assess the reliability and validity of the test separately for each group and to demonstrate the comparability of the obtained values. In this connection, it is particularly critical that comparability of construct validity for the different groups be appraised; this can be done by examining the patterns of correlates of the test with other measures to see if they are similar across groups. In addition, if the test is to be used for purposes of selection, classification, or guidance, its predictive validity should also be separately evaluated where technically feasible, taking care to check that there is a common criterion uniformly applied across groups.

2. The Test May Involve Irrelevant Difficulty

The estimates of the capabilities of minority/poverty groups

derived from certain tests may be systematically lower than they should be because of irrelevant difficulties in the testing situation. This kind of distortion represents a bias in the measurement or estimation of ability levels in the same sense that a sample statistic which uniformly deviates from a population parameter is a biased estimate.

Some examples of irrelevant difficulty are a test format requiring a child to read the instructions for a task intended to assess listening comprehension ability, an answer marking procedure that is almost as difficult as the problems posed by the test itself, and a time limit that is severely restrictive when the testing task requires varying amounts of reflection by the respondents.

Other potential sources of irrelevant difficulty include:

(a) Items that are more germane to one group than to another. One way of uncovering such items is to search for response distributions that exhibit item-by-group interactions, thereby revealing items that are relatively more difficult or relatively easier than the majority of the items for one group as opposed to another. Items differentially favoring males or females have been uncovered over the years and their distinctive properties elucidated (Coffman, 1961), but relatively few studies have addressed themselves to the identification of items that differentially favor one racial or ethnic group over another in this sense (Cleary and Hilton, 1968). Whenever possible it would be a desirable addition to standard item-analysis practice

to search for such items routinely, as well as to examine the possibility of the differential attractiveness of multiple-choice distracters to different population groups. Such investigations would increase our understanding of differential item properties and of individual and group differences in item response and would provide an additional empirical basis for judging how appropriate a test is for particular individuals and groups.

In the past, when occasional biased items were uncovered on a test, their appearance was usually defended on the grounds of their small contribution to total test variance or on the basis of the inclusion of a sufficient number of counter-biased items to balance their influence. These arguments addressed themselves to the problem of biased items as threats to validity but not to the social and educational consequences of administering biased items to individuals they are biased against.

(b) Testing conditions that make some individuals feel anxious, threatened, or alienated. The adverse consequences of such negative affects on test performance are potentially quite serious (Katz, 1970), and vigorous attempts should be made to prevent their occurrence, through such steps as utilizing familiar and congenial settings and administrators, reducing the adversarial quality of the testing situation, emphasizing to the examinee the positive values of the information being collected for educational and developmental purposes, etc.

(c) Differences in test wiseness. Individuals and groups differ in their degree of test wiseness and in their familarity with various test-taking strategies, and this inexperience with effective approaches to test taking may place some at a disadvantage, at least initially. This is likely to be a more serious problem with young children than with high school or college students. The differential effects of variations in test wiseness may be reduced by the use of clear and detailed instructions and by exposure to practice items, preferably with feedback. In addition, test-taking strategies can be taught. To the extent that such strategies of test taking are also strategies of thinking and problem solving, this effort would be generally beneficial to the student, and it would tend to increase not only the test scores but also the intrinsic validity of the test (Gulliksen, 1950). To the extent that the test-taking strategies are primarily adaptive to particular properties of the test design, the time might be better spent in improving the tests and the techniques of administering them.

3. The Test May Accurately Reflect Ability or Achievement Levels

Low scores per se do not necessarily indicate bias in measurement. Many of the abilities assessed by conventional tests develop out of educational, social, and family experiences over many years. Low test scores may represent an unbiased assessment of ability levels that have been limited by the cumulative impact of poverty, prejudice,

inequality of educational opportunity, and other factors. The ghetto child, for example, who has attended a succession of inferior schools (with all that "inferior" implies for the quality of teachers, instruction, and facilities) and comes from a home where books and other learning supports have always been in short supply cannot be expected suddenly to handle with competency the materials and problems of the "average" curriculum--or standardized test. The bias under these conditions is not in the estimation of the ability levels but in the social forces that inhibited development. Test scores in these circumstances then become a powerful monitor of the inequities of the educational and social system, as well as a blueprint for constructive educational action at the individual level.

## The Appropriateness of Test Use and the Question of Fairness

The underestimation of ability and achievement levels does not necessarily imply unfairness in the use of those scores. Bias in measurement and unfairness in practice are often concomitants, to be sure, but they do not have to go hand in hand. Consider a selection or placement situation where a test is valid for two groups but one group characteristically obtains lower scores than the other because of irrelevant difficulty or other sources of bias; assume also that there is not a corresponding difference in criterion performance. If the selection or placement decision is made in terms of acceptable levels of predicted criterion performance, then separate

cut-off scores or regression functions would (and should) be utilized for the two groups, with an appropriately lower cut-off being employed in the low-scoring group. This is tantamount to adding points to the scores of the low-scoring group--not as a general strategy as described by Williams (1970b), but under certain specific circumstances where the procedure is clearly justified; i.e., when there is external evidence of measurement bias and sufficient information to estimate the size of the effects for a particular purpose. In this example, then, we have assumed a systematic underestimation of ability levels in one group due to measurement bias and have illustrated the possibility of utilizing within-group test validity to produce a selection procedure that is not unfair to the low-scoring group.

This example dramatizes the possibility that a test might have a different validity coefficient or a different regression function for a minority/poverty group than for a middle class group and that the general use of prediction equations derived from the White majority might unfairly penalize minority individuals in selection or placement situations. Since such an eventuality cannot be discounted on logical grounds, testing practitioners should be constantly alert to the prospect. However, investigations thus far have not produced many examples of this kind of unfairness in educational settings. On the contrary, although there may sometimes be group differences in validity coefficients and regression lines, when predictions of

academic performance are based upon regression equations suitable for the majority group, then minority individuals are predicted to do about as well or somewhat better than they actually do (Cleary, 1968; Kendrick & Thomas, 1970; Stanley & Porter, 1967; Temp, 1970).

Importance and Relevance

Central to the issue of fairness in the applications of testing is the question of the appropriateness of the selected test for the proposed purpose. Judgments of whether or not a specific test should be used for a particular purpose must take into account the relevance of the attributes measured to the intended criterion and the importance of the information obtained for the given objective (APA, 1969; APA, 1970). The appropriateness of the objective itself should also be evaluated, and this places us squarely in the arena of social values and public responsibility.

Although judgments concerning test use must be relative to specific purposes (for a test may be fine for one purpose and terrible for another), the decision should also take into account the possibility of additional consequences or side effects attendant upon the use. In addition, the potentiality for misuse must also be considered, as well as the possibility of safeguards to protect against it.

Side Effects of Testing

It should be recognized that the administration of a test may

have many consequences in addition to the intended assessment. These may be either positive or negative and may affect both the examinee and the examiner.

To begin with, the taking of tests can be a rewarding enterprise and may even be fun. Our experiences in testing over a thousand Black preschoolers in the ETS Longitudinal Study of Disadvantaged Children and Their First School Experiences (Anderson, 1970; ETS, 1969) have clearly demonstrated that this is possible. At a minimum, a test should be constructed to provide a pleasant or challenging experience for a child. At best, it would also have some instructional value in its own right. The measures developed in the ETS "Let's Look at First Graders" series (1965), for example, appear to have this property. For older students, who have more understanding of what testing is about, a good test can also serve to define the objectives of a course of study, to highlight important concepts, and to stimulate the synthesis of ideas.

On the negative side, there are numerous examples of test taking as a frustrating experience for students: when the relevance of many of the items is unclear, when content is ambiguous or inaccurate, when tasks are at an inappropriate difficulty level, when--in the case of "school" tests--the test is not closely related to the student's study assignment, when there are irrelevant difficulties, when the test reinforces negative feelings the student already has toward the educational system. In addition, the context

of testing can supply its own negative affect. Not many students can be expected to enjoy taking a test that may eliminate them from a competition (in contrast, for example, to one focusing less on selection and more on diagnosis for training or self-improvement).

The test administrator can also be affected by tests. If he views tests as an imposition--and many teachers do, especially when they had little voice in the decision to use them--he may not only conduct an unprofessional administration but also communicate his feelings in subtle or unsubtle ways to the examinees, with obvious consequences for their performance. Or, if he is not tuned in to the purposes of testing, he may engage in such lamentable practices as teaching the test items specifically, with little regard for the more general processes involved.

On the other hand, some testing sessions--particularly those conducted individually or in small groups as is most appropriate for younger children--provide an excellent opportunity for a teacher to observe a child intensively, to study his reactions and coping behaviors, and to identify types of situations that disturb him. These observations, made in a fairly standard situation affording comparability across many students, may provide far more valuable information than scores. In addition, a good assessment battery can do much to promote consideration of the complexity of students and the broad range of skills, attitudes, achievements, social competencies, etc., that characterize their development and underlie their responses to educational and social stimuli. A battery

appropriate for use at the local level but developed nationally can also provide teachers with some protection against insularity. It can remind them of the broad goals of education and call attention to performances that other educators value.

It is important to recognize that such effects on the tested and the testers can occur relatively independently of the interpretation of results. Some of the arguments of the Black Psychologists can be seen as referring to the test-taking process and the side effects it may have. It is the responsibility of test developers, selecters, and users, first, to recognize the possibility of such effects and, second, to edit, evaluate, and use their tests in such a way that they foster only the positive ones.

## The Limits of Test Use

Most of the possible misuses of tests can be attenuated if appropriate safe-guards and guiding principles are adopted in advance. In this regard, two critical problems warrant special comment; namely, the problem of misinterpretation and the problem of secondary use of test results.

1. The problem of misinterpretation. A large portion of the problem of poor interpretation is really a problem of poor thinking. It is misinterpretation based upon a misconception of the phenomenon being measured or an exaggerated expectation about the infallibility of tests. One form of misconception that is particularly wide-spread

is the presumption that test scores reflect fixed levels of capacity. Another type of frequent misinterpretation derives from the tendency to take seriously insignificant differences between scores. The former error can only be overcome by training and enlightenment, but the latter error of overinterpreting small differences can be substantially reduced by judicious presentation of results--through the use of percentile bands, stanine scores, quartiles, or other devices. In general, a number of kinds of misinterpretation can be avoided by careful presentation of scores, and one important guiding principle is that the presentation should relate as directly as possible to the types of decisions to be made on the basis of the results.

For example:

| | |
|---|---|
| For student self-guidance into various educational programs or courses-- | probability tables, showing predictions of future success and/or satisfaction associated with his present test performance |
| For teacher selection of instructional activities for students-- | individual student profiles, with level of performance in each area linked to appropriate curricula and materials |
| For superintendent evaluation of the effectiveness of innovative programs in his system-- | school-by-school distributions of criterion test results, as related to such factors as "input" level of students and program intensity |
| For curriculum specialists' revisions-- | summaries of results (including errors), classified by relevant portions and goals of the curriculum and student characteristics (how did the curriculum work for whom?) |

2. The problem of secondary use. The secondary use of test results raises a key issue in the ethics of testing. When, if ever, is it appropriate to use test results collected at one point in time for one purpose at another point in time for either the same or a different purpose? Suppose someone wanted to use third-grade test scores in counseling 11th graders about their chances of success on a college admissions test. Or information from a biographical inventory collected to determine a school district's eligibility for Title I funds to place students in remedial classes. Or 9th grade biology scores to support a recommendation about a student's eligibility for advanced placement in biology. Or test data from the 1947 freshman class to write a paper on how freshmen have changed in the last quarter century.

The decision on secondary use must face the same two requirements as the decision on primary use--the need to justify the proposed procedure on scientific grounds and in terms of its potential social consequences. In terms of scientific criteria alone, such actions as those just described might be justified if it could be shown empirically that the test results were indeed valid predictors of performance at another point in time or in another arena--or, in the case of the last example above, that the design of the analysis and the measurement properties of the instruments allowed valid comparisons.

But the ethical question of "Should these actions be taken?" cannot be answered by a simple appeal to empirical validity alone. The various social consequences of these actions must be contended with, especially those bearing on issues of invasion of privacy, confidentiality of records, and client welfare. As might be expected from the diversity of potential secondary uses illustrated above, there is no single principle or set of principles that can appropriately be applied to all situations, types of measurement, or population groups. Value judgments have to be made about each, taking into account local attitudes and personnel involved as well as more general scientific and humanistic concerns.

The conflict between advancement of science and the general social welfare on the one hand and protection of the rights and privacy of individuals on the other is especially salient when test results are proposed for some use other than that originally intended. For example, the decision of a few years ago to remove all ethnic identification from student and other personnel records was viewed by most laymen as a guarantee of civil rights but by many social scientists as an unfortunate and needless constraint, for it prohibited them in many instances from developing some of the very information that government and educational agencies and representatives of minority groups are now clamoring for--information that would enable better placement of minority group members in training programs and jobs or would throw light on

possible discriminatory practices.

## The Social Consequences of Not Testing

In spite of imperfections in current tests and testing practices, it is clear that educational and psychological tests serve many critical functions--not always optimally, to be sure, but better than proposed alternatives. As we face the recent call by the Association of Black Psychologists for an immediate moratorium on all testing of Black people, we must pause to ponder what might be lost by the elimination of testing.

To begin with, the needs that testing serves would still exist and would be addressed by other means. If objective and standardized tests were not available, people would revert to the uses of the past --to subjective appraisals such as the interview and inquiries into ancestry. And one consequence of this seems clear--the likelihood of bias and discrimination would increase. A recent study (Sparks & Manese, 1970) has shown, for example, that minority group members are rated systematically lower on pre-employment interview dimensions in the absence of interviewer knowledge of test scores, as compared with the level of interview ratings obtained for comparable groups when interviewers have such knowledge. Other research (Flaugher et al., 1969) indicates that supervisor's ratings vary as a function of the race of the rater and the race of the ratee, as does the validity of predicting such ratings as criteria of job performance.

In addition, without tests in educational and job-training programs, teachers and counselors would be forced to rely only upon observations of skills and deficiencies during the course of the program. Although this might in many instances provide important information upon which to base subsequent treatment, it would also require a prolonged period of time to collect and would rarely be done systematically. Practitioners would thus be faced with the prospect of slow assessment, whereby valuable educational time must be diverted to preliminary observation before specialized treatment can be sensibly applied.

The elimination of tests would also mean the loss of one of the best ways for teachers to acquire a useful appreciation of the broad range of competencies and traits that characterize human behavior or to develop needed sensitivities to the nuances of cognitive growth. An increased parochialism might spread throughout education because of the absence of a national normative perspective and the limitation of access to concrete examples of what other educators deem important to assess. And of utmost importance, there would be an absence of yardsticks for gauging the effectiveness of educational programs and for evaluating the equity of the educational system.

Thus, the social consequences of not testing are extreme. Tests may be eliminated only at a cost, and a large portion of that cost would be increases in discrimination and ignorance.

## References


American Psychological Association-Job testing and the disadvantaged. APA Task Force on Employment Testing of Minority Groups. American Psychologist, July 1969, 24, 637-650.

American Psychological Association-Psychological assessment and public policy. American Psychologist, 1970, 25, 264-266.

Anderson, S. B. From textbooks to reality: social researchers face the facts of life in the world of the disadvantaged. In Jerome Hellmuth (Ed.), Disadvantaged Child Volume 3. Compensatory Education: A National Debate. New York: Brunner/Mazel Publishers, 1970, 226-237.

Brown, H. R. Die nigger die! New York: The Dial Press, Inc., 1969.

Cleary, T. A. Test bias: prediction of grades of negro and white students in integrated colleges. Journal of Educational Measurement, Summer 1969, 5, 115-124.

Cleary, T. A., & Hilton, T. L. An investigation of item bias. Educational and Psychological Measurement, 1968, 28, 61-75.

Coffman, W. E. Sex differences in response to items in an aptitude test. 18th Yearbook, National Council on Measurement in Education, 1961, 117-124.

Disadvantaged children and their first school experiences: ETS-Head Start Longitudinal Study--From Theory to Operations. Project Report 69-12, Educational Testing Service, August 1969, Princeton, N. J.

Flaugher, R. L., Campbell, J. T., & Pike, L. W. Prediction of Job Performance for Negro and White Medical Technicians. Project Report 69-5, Educational Testing Service, April 1969, Princeton, N. J.

Gulliksen, H. Intrinsic validity. American Psychologist, 1950, 5, 511-517.

Jackson, D. N., & Messick, S. The ethics of assessment. In D. N. Jackson & S. Messick (Eds.), Problems in Human Assessment. New York: McGraw-Hill, Inc., 1967.

Katz, I. Experimental studies of negro-white relationships. In L. Berkowitz (Ed.), Advances in Experimental Social Psychology, 5 New York: Academic Press, in press.

Kendrick, S. A., & Thomas, C. L. Transition from school to college. Review of Educational Research, 1970, 40, 151-179.

Let's look at children. Educational Testing Service, 1965, Princeton, N. J.

Messick, S. Personality measurement and the ethics of assessment. American Psychologist, 1965, 20, 136-142.

Sparks, C. P., & Manese, W. R. Interview ratings with and without knowledge of pre-employment test scores. The Experimental Publication System, 1970, 1-10.

Stanley, J. C., & Porter, A. C. Correlation of scholastic aptitude test scores with college grades for negroes versus whites. Journal of Educational Measurement, Winter 1967, 4, 199-218.

Temp, G. An examination of test bias: some data, some questions. Educational Testing Service, 1970. Research Bulletin, in press.

Williams, R. L. Black pride, academic relevance, and individual achievement. The Counseling Psychologist, 1970a, 2, 18-22.

Williams, R. L. From dehumanization to black intellectual genocide: a rejoinder. Clinical Child Psychology Newsletter, Fall 1970b, 9, 6-7.